Hammill Institute on Disabilities

Topics in Early Childhood Special Education
2017, Vol. 37(3) 164–175

DOI: 10.1177/0271121416669924
tecse.sagepub.com
SAGE

Using Generalizability Theory to Examine the Dependability of Scores From the Learning Target Rating Scale

Tara W. McLaughlin, PhD[1,2], Patricia A. Snyder, PhD[1], and James Algina, EdD[1]

Abstract

The Learning Target Rating Scale (LTRS) is a measure designed to evaluate the quality of teacher-developed learning targets for embedded instruction for early learning. In the present study, we examined the measurement dependability of LTRS scores by conducting a generalizability study (G-study). We used a partially nested, three-facet model to estimate the variance of LTRS scores due to teachers, children, learning targets, and raters. We used these variance components to conduct a decision study (D-study) to investigate how data collection and scoring design affected the dependability of scores and to help inform future use. Findings supported the dependability of LTRS scores when used with one rater and when at least six learning targets are obtained from teachers. We discuss potential refinements for the LTRS based on the G- and D-studies and implications for using it in practice.

Keywords

embedded instruction, measurement, learning targets, generalizability study

Current recommended practices in early intervention and early childhood special education emphasize the need for intentional and systematic instruction to promote child participation, engagement, and learning in inclusive early learning settings (Division for Early Childhood [DEC], 2014). To provide effective targeted or individualized instruction for young children with disabilities in these settings, practitioners need instructional goals that specify what is to be taught, when and where instruction will occur, how instruction might occur, and how the effectiveness of instruction will be evaluated (Snyder, Hemmeter, Sandall, McLean, & McLaughlin, 2013). Embedded instruction is an evidence-based approach that emphasizes the *what to teach*, *when to teach*, *how to teach* and *how to evaluate* components of effective instruction. Embedded instruction has been used to increase access to the general preschool curriculum, enhance child engagement, support children's social interactions, and improve learning outcomes for young children with disabilities (Horn, Lieber, Li, Sandall, & Schwartz, 2000; McLaughlin & Snyder, 2014; Snyder et al., 2015).

Embedded instruction in early childhood special education has evolved from a history of practices focused on naturalistic teaching, including incidental teaching and milieu teaching (Snyder et al., 2015). For young children, a primary focus of embedded instruction is on contextualized learning in developmentally appropriate activities. Children learn skills that are aligned with their individualized educational programs (IEPs) and support their participation, independence, or interactions with adults and peers in these activities. In early childhood, the instructional environment (i.e., when to teach) is focused on activity-based and routine-based contexts and the instructional content (i.e., what to teach) is focused on behaviors or skills for active participation and engagement in these settings that are aligned with early learning guidelines or standards to ensure children's access to the general preschool curriculum. These features, in turn, influence the way in which embedded instruction is implemented and evaluated (i.e., how to teach and how to evaluate).

To support the quality of instructional practices and measure effective teaching, observational and judgment-based rating instruments are being adopted in early care and education programs and used in intervention research at a rapid pace (cf. Zaslow et al., 2011). Semmelroth and Johnson (2014) noted that efforts focused on developing and applying these types of measures of effective teaching have often

[1]University of Florida, Gainesville, USA
[2]Massey University, Palmerston North, New Zealand

Corresponding Author:
Tara W. McLaughlin, Institute of Education, College of Humanities and Social Sciences, Massey University, Private Bag 11-222, Palmerston North, New Zealand.
Email: t.w.mclaughlin@massey.ac.nz

omitted specific learning settings and subgroups of students with diverse needs. Given the instructional and interactional differences across the learning life span (infant/toddler, preschool, elementary, secondary), there is a need for related but context-specific measures of effective instructional and interactional practices. For example, measures designed for promoting effective or recommended interactional or instructional practices for preschool-age children with disabilities might be an important focus for the field of early childhood special education.

To measure preschool teachers' implementation of embedded instruction, Snyder et al., 2011 designed a suite of measures to examine key components and outcomes of embedded instruction focusing on what, when, and how to teach. The present study focused on examining the dependability of scores for the measure designed to focus on the quality of instructional goals written by preschool teachers as part of the what to teach component of embedded instruction.

The what to teach component involves preschool teachers writing high-quality embedded instruction goals, based on child learning priorities, to inform the design, delivery, and evaluation of embedded instruction for early learning. Written statements of children's learning priorities help to ensure accountability and support instructional decision-making and progress monitoring (Drasgow, Yell, & Robinson, 2001; Goodman & Bond, 1993; Huefner, 2000). Yet, research has shown many practitioners have difficulty identifying important skills for child learning and writing learning priorities to guide instruction, particularly instruction provided in the context of the general preschool curriculum (Giangreco, Dennis, Edelman, & Cloninger, 1994; Grisham-Brown & Hemmeter, 1998). For example, more than 30 years of research has shown that written IEP goals or objectives often do not reflect key aspects identified as necessary to meaningfully inform instruction and progress monitoring (Christle & Yell, 2010). Issues have included (a) lack of generalized outcomes or age-appropriate content (Lynch & Beare, 1990), (b) insufficient emphasis on active participation (Giangreco et al., 1994), (c) skills that are not measurable or do not include performance criteria (Pretti-Frontczak & Bricker, 2000), (d) failure to link skill selection to assessment data (Giangreco et al., 1994), and (e) skills focused on limited domains (e.g., preacademic skills; Michnowicz, McConnell, Peterson, & Odom, 1995).

The *Learning Target Rating Scale* (LTRS; Snyder et al., 2009) is an investigator-developed measure designed to quantify the quality of teacher-developed instructional goals for embedded instruction for early learning. The LTRS was developed for use in a research study; yet, it could be useful as a guide for preschool teachers and preservice teachers using or learning about embedded instruction. In the present study, we examined the measurement dependability of LTRS scores, obtained under conditions of intervention research, using a generalizability theory (G-theory) approach. We conducted a generalizability study (G-study) and a decision study (D-study) to address three research questions:

Research Question 1: How much variance in observed teacher LTRS total scores is associated with raters, children, and learning targets?
Research Question 2: How much variance in observed teacher LTRS dimension scores is associated with raters, children, and learning targets?
Research Question 3: Varying the number of raters, children, and learning targets, what study-designs produce dependable teacher LTRS total scores?

G-theory allows researchers to explore multiple sources of variance in scores simultaneously and can be used to forecast dependability of scores under different conditions (Shavelson & Webb, 1991). This approach to examining score dependability is particularly useful to understand influences affecting the development and use of observational measures. Information obtained is also informative for teachers and those who support or evaluate teachers' practices to understand what is being measured and how it is being measured as these relate to obtained *quality* scores.

Method

Development of the LTRS

Development of the LTRS was guided by recommended practices for scale construction (Crocker & Algina, 2006). This included clarification of key concepts, a literature review, adaptations from two existing measures (i.e., Hunt, Goetz, & Anderson 1986; Notari & Bricker, 1990), expert review, and a comprehensive pilot of the Learning Target Rating Scale Research Version 1.0 (LTRS-V1; Snyder et al., 2008).

The LTRS focuses on embedded instruction goals known as *learning targets*. Learning targets refer to written statements that specify an embedded instruction learning priority for a young child. A learning target should be an immediate instructional priority intended to guide instruction on a day-to-day basis to support the child's engagement and learning and the skill specified in the learning target should align with early learning standards, the preschool curriculum for all children, and a child's annual IEP goals or objectives.

The LTRS-V1 was piloted with a sample of learning targets obtained from 13 preschool teachers who were using an embedded instruction approach. Procedures for training, administration, and scoring were established during the pilot. The LTRS-V1 was revised based on pilot data, including interrater agreement and internal consistency score reliability, rater feedback, teacher feedback, and expert review.

Indicators with low interrater agreement or poor internal consistency were revised and decision rules for scoring were added. Additional dimensions and indicators were also developed. For example, we added a dimension to delineate the steps to specify an immediate instructional priority for a child. Following the revision process, the LTRS Research Version 2.0 (LTRS-V2; Snyder et al., 2009) rating scale and manual were finalized. This latter version of the LTRS was used in the present G-study.

Sample of Learning Targets

The sample of learning targets for the present study came from a randomized controlled potential efficacy trial of a professional development (PD) intervention focused on supporting preschool teachers' use of embedded instruction, which involved 36 teachers and 106 children and was conducted over the course of a preschool year (Snyder et al., 2016). Teachers contributed four learning targets for two or three children in their classroom who were study participants. Learning targets were collected on five occasions, but, due to resources available, only targets from three occasions were used in the present study. The three occasions were baseline, before intervention was provided to teachers; halfway through the study; and toward the end of the study. The total number of targets for the present study was 1,237 with 420 at the first occasion, 422 at the second occasion, and 395 at the third occasion.

Participants

The teachers who contributed learning targets were from three school districts located in separate regions of the country: Northwest (n = 11), Midwest (n = 12), and Southeast (n = 13). The majority of preschool classrooms were located on local elementary school campuses. Thirty-five of the 36 teachers were female and all were certified. On average, teachers had 7 years of experience in early childhood settings. All participating children had an IEP, and their mean age was 40.4 months (SD = 8.7). The majority of participating children were male (77%) and were eligible for special education services based on having a developmental delay (59%).

Measure

The LTRS-V2 (Snyder et al., 2009) is a summated judgment-based rating scale and includes 16 indicators organized under six dimensions: (a) behavior statement (v = 4), (b) age appropriateness (v = 2), (c) functionality (v = 3), (d) generality (v = 2), (e) instructional context (v = 2), and (f) measurability (v = 3). Table 1 shows the dimensions, associated indicators, and a brief definition for the indicators. The written learning target is scored dichotomously for each indicator (0 = *does not meet indicator* or 1 = *meets indicator*). There are two indicators for which a *not applicable* (N/A) rating can be applied (see Table 1). The measure is designed to provide an overall rating of learning target quality across the dimensions and indicators. The measure is not designed to evaluate whether the content of the learning target is appropriately individualized for a specific child, as this would require in-depth knowledge about individual children that is unavailable to raters.

Trained raters use the LTRS manual to evaluate each indicator for each learning target. The manual includes a definition and description of each indicator, with related examples, clarifying questions, and decision rules to guide ratings. A score can be computed for each learning target; however, the measure is intended to produce an overall teacher score with measures of central tendency and variability available for each teacher's learning targets.

Procedures

Learning targets. Teachers received a form on which to write learning targets. The form indicated,

> learning targets describe the skills or behaviors that you are working on with the target child in your classroom now and over the next few weeks. Learning targets might be taken from IEP goals and objectives or you might write behavioral statements for other skills that you are focusing on with the child.

Teachers were asked to write four learning targets they were currently working on for each child in their classroom enrolled in the study. On all occasions, the same LTRS administration procedures were used for teachers in control and intervention conditions. However, subsequent to the first occasion, teachers in the intervention condition received workshops, resource guides, and a form of coaching as part of a PD intervention. The PD intervention was designed to help teachers learn about embedded instruction, including how to write quality learning targets.

Rater training. Two raters completed training and served as raters for the present G-study. Training included reading the LTRS manual, reviewing the scoring sheet, reading examples and non-examples for each indicator, completing ratings for practice sets of learning targets and comparing ratings with an expert standard, and receiving feedback from the first and second authors of the LTRS. Following training, each rater had to reach at least 80% agreement with an expert standard across four practices sets of learning targets before coding for the present study.

LTRS scoring. Each trained rater scored each learning target for each teacher from each of the three measurement occasions. Interrater agreement was checked after each rater rated

Table 1. LTRS-V2 Dimensions, Indicators, and Definitions.

Dimension	Indicator	Definition
Behavior statement	Behavior specified	The behavior stated in the learning target specifies an action the child should do.
	Demonstration specified	A statement of how the child should demonstrate the behavior is included in the learning target or is synonymous with the behavior specified.
	Exemplars provided	The learning target specifies two or more exemplars of what the skill should look like.
	Written clearly	The learning target is written in clear, jargon-free language.
Age appropriate	Materials	Materials specified in the learning target are developmentally appropriate for same-age peers without disabilities. (N/A possible)
	Skill	The skill specified in the learning target is appropriate for same-age peers without disabilities.
Functionality	Access to preschool curriculum	The skill specified in the learning target is related to developmental domains or content associated with general early childhood curricula; early learning guidelines, standards, or benchmarks.
	Independence	The skill specified in the learning target is a critical skill for activities of daily living or a skill that supports participation in activities.
	Interactions with peers	The learning target includes a statement that the child should do something with a peer with or without disabilities. (N/A possible)
Generality	Multiple forms	The way the skill is written provides opportunities for the child to show different forms of the behavior and achieve the same function.
	"Embeddable"	The skill specified in the learning target can occur two or more times in the context of classroom activities, routines, or transitions.
Instructional context	Natural environments	The learning target is specified in a way that implies the skill will be taught in natural environments.
	Across activities	The learning target includes a statement that specifies the skill should occur across activities, routines, or transitions.
Measurability	Observable	The skill is specified so it can be counted or measured.
	Conditions	The learning target is written so the conditions under which the skill is to occur are specified.
	Criterion stated	The learning target includes a statement of criterion.

Note. LTRS-V2 = Learning Target Rating Scale Research Version 2.0.

200 learning targets. Raters were asked to review the manual for indicators where agreement levels were lower than 80% before proceeding with rating additional learning targets. Raters were blind to the experimental study conditions of the participating teachers and children. Raters entered scores directly into an Excel™-based scoring sheet. LTRS scores were calculated as the number of indicators rated as present divided by 16 minus the number indicators rated as not applicable. This proportion was then multiplied by 100 to generate a percent present score for all indicators and each dimension. Total teacher scores are percentage of indicators present scores averaged across learning targets and children.

Data Analysis

We used G-theory (Shavelson & Webb, 1991; Webb, Shavelson, & Haertel, 2007) to examine the measurement dependability of LTRS scores. G-theory was selected to investigate the dependability of LTRS scores over other methods available through classical test theory because the sources of variance in scores can be identified and used to examine multiple measurement conditions (Thompson, 2003).

In classical test theory, the observed scores are composed of true and error scores ($X = T + E$), with the error term undifferentiated (Crocker & Algina, 2006). Using various reliability coefficients from classical test theory, sources of error (e.g., raters, occasion, form) can be examined one at a time. In G-theory, variances due to multiple sources of error are examined simultaneously, isolated, and used to further examine the dependability of scores under different measurement conditions (Crocker & Algina, 2006; Shavelson & Webb, 1991). G-theory is a generalization of classical test theory in which the components of observed score variances are estimated from the object of measurement (e.g., the person for whom measurements are being made), the facets of measurement (e.g., sources of error due to measurement features or conditions), and their combination. Examining the dependability of test scores is a generalization of examining the reliability of test scores in classical test theory (Webb et al., 2007). In G-theory, the variance components that affect the dependability of generalizations about universe scores from observed scores (i.e., a generalization about the true score from classical test theory) are estimated in a G-study. The variance components

can be used to forecast the dependability of scores obtained under different designs in the future (i.e., a D-study; Shavelson & Webb, 1991). These are reported as generalizability (G) coefficients and are used to assess dependability. A D-study is a generalization of using the Spearman–Brown prophecy formula, which is used in classical test theory to forecast the effect of test length on reliability.

Design. We used a partially nested, three-facet model to conduct the G-study using LTRS total scores and dimension scores. Learning targets were nested within children and children were nested within teachers, while raters were fully crossed with learning targets, children, and teachers. Teachers, children, learning targets, and raters were included in the model as random effects. Our G-study design, with raters crossed with targets, children, and teachers allows estimation of G coefficients for a D-study design in which the same crossing is included as well as for a D-study design in which raters are nested in targets, children, and teachers. As noted previously, the learning targets were collected as part of a potential efficacy trial of a PD intervention; thus, we expected the quality of learning targets, and therefore LTRS scores, to improve over time for teachers in the intervention condition. Consequently, in our G- and D-studies, we were not interested in how well LTRS scores generalized over time and we chose to analyze the three occasions separately.

G-study. The aim of a G-study is to estimate the variance components based on available data. PROC VARCOMP in SAS 9.3 was used to estimate a random effects model for each of the three occasions. For total scores from the LTRS, we estimated the percentage of variance accounted for by the sources of variance in the model for each of the three occasions. For dimension scores, we estimated the percentage of variance for Occasion 2. Occasion 2 was selected for dimension scores because it exemplifies a time point at which teachers' score variance had increased relative to variance before the embedded instruction PD intervention was implemented, but was not expected to be overly influenced by participation in PD intervention conditions as in Occasion 3.

D-study. In a D-study, variance components estimates are used to calculate G coefficients for potential future designs. These designs can vary in terms of the number of levels for the facets of measurement (i.e., the sources of measurement error). In the present study, the facets of measurement were raters, learning targets, and children. The object of measurement is the source of universe score variance (a generalization of the true score variance in classical test theory). Teachers were the object of measurement because the purpose of the LTRS is to provide total scores for teachers.

In the D-study, we used the variance components estimates to calculate G coefficients for relative (norm-referenced) decisions about teachers' scores. We calculated G coefficients for two possible designs: Design 1 in which raters are fully crossed with teachers, that is, targets written by each teacher are rated by each rater and Design 2 in which raters are nested in teachers, that is, targets written by each teacher are rated by a different set of raters. To examine the dependability of scores on each occasion, we calculated G coefficients for designs with one and two raters, three children per teacher, and four leaning targets per child. The aim of a D-study is to use the variance components to design efficient measurement conditions for future use. For each design, we used the G-study results for Occasion 2 to calculate G coefficients under 49 different measurement conditions by varying the number of raters, learning targets, and children to determine adequate conditions for total score dependability.

Results

G-Study

The percentage of variance of total LTRS scores accounted for by the different sources of variance in the model was similar across the three occasions, with learning targets and teachers accounting for the largest percentage of variance (see Table 2), while raters and children accounted for a small percentage of variance. The percentage of error variance decreased over the three occasions. Across dimension scores (see Table 2; shown for Occasion 2 only), learning targets accounted for the largest percentage of variance, except for the instructional context and measurability dimensions in which teacher accounted for more variance. The percentage of error variance varied across the dimension scores.

D-Study

Design effects. Consistent with the goal of D-studies to estimate G coefficients for a variety of designs that might be used in the future, G coefficients for total scores and dimension scores are shown in Table 3 for two potential D-study designs. For Design 1, G coefficients for total scores were moderate to high, ranging from .77 to .90 under conditions in which raters would be crossed with teachers. G coefficients ranged from .74 to .90 if raters were to be nested in teachers (Design 2).

The functional and generalizability dimensions had the lowest dependability (.70–.76) and the measurement dimension had the highest dependability (.91–93) across the designs evaluated. We examined these two designs because in a field-based intervention research study, the same raters might rate targets from subsets of teachers, but raters might not be fully crossed with teachers. Thus, the conditions of future studies might have some raters crossed with teachers and others raters nested in teachers. Given a combination of

Table 2. Percentage of Total Score and Dimension Score Variance by Source of Variance.

	Total scores occasion			Dimension scores Occasion 2					
Source of variance	1	2	3	BS	AA	F	G	IC	M
Teacher	**24.87**	**43.78**	**33.87**	**30.03**	**29.70**	**18.03**	**20.98**	**46.27**	**59.86**
Child	2.99	0.94	1.78	3.70	0.84	0.00	0.00	0.00	3.59
Learning target: Child: Teacher	**53.97**	**43.13**	**56.63**	**49.22**	**52.08**	**59.30**	**58.14**	**29.43**	**26.55**
Rater	1.31	0.36	0.03	0.30	0.00	0.28	0.36	0.21	0.00
Teacher by rater	0.48	0.83	0.00	1.44	0.13	0.62	2.34	1.24	1.46
Child: Teacher by rater	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.48
Error	**16.39**	**10.95**	**7.69**	**15.31**	**17.25**	**21.77**	**18.19**	**22.85**	**8.05**
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Note. Bolded results are used to draw attention to the largest sources of variance across facets. BS = behavior statement; AA = age appropriate; F = functionality; G = generality; IC = instructional context; M = measurability.

Table 3. D-Study Designs: G Coefficients for Total Scores and Dimension Scores.

	Generalizability (G) coefficients								
	Total scores occasion			Dimension scores Occasion 2					
Study conditions	1	2	3	BS	AA	F	G	IC	M
Design 1: Raters crossed with teachers									
Three children, 4 targets, and 1 rater	0.77	0.89	0.85	0.79	0.83	0.71	0.71	0.89	0.91
Three children, 4 targets, and 2 raters	0.80	0.90	0.86	0.82	0.85	0.75	0.76	0.92	0.93
Design 2: Raters nested in teachers									
Three children, 4 targets, and 1 rater	0.74	0.88	0.85	0.78	0.83	0.70	0.70	0.89	0.91
Three children, 4 targets, and 2 raters	0.78	0.90	0.86	0.81	0.85	0.74	0.75	0.92	0.93

Note. D-Study = decision study; BS = behavior statement; AA = age appropriate; F = functionality; G = generality; IC = instructional context; M = measurability.

these designs is likely, a conservative approach is to report the G coefficient for Design 2.

Facet effects. To investigate the effect of the number of children, learning targets, and raters on the G coefficients, we calculated G coefficients for various combinations of numbers of children, learning targets, and raters. Adding raters had minimal impact on the dependability of scores (i.e., less than .03 difference in G coefficients across all conditions). The small effect of number of raters was due to the small percentage of variance due to the main effect of rater and its interactions with teachers and targets (see Table 2). Increasing the number of learning targets or increasing the number of children, however, had a notable effect on the dependability of scores. Table 4 shows the G coefficients when single raters are nested in teachers (the most conservative coefficient) for different configurations of learning targets and children. Assuming an adequate G coefficient is 0.80, this can be achieved by different combinations of numbers of children and learning targets that result in a minimum of six learning targets written by the teacher. As noted earlier, the variance component for children was smaller than the variance component for learning targets. Nevertheless, increasing the number of children has a substantial effect on the G coefficient. This is because the increase in the number of children causes an indirect increase in the total number of

Table 4. D-Study: G Coefficients for Total Scores Varying the Number of Children and LTs.

	Number of children						
Number of LTs	1	2	3	4	5	6	7
1 LT	0.44	0.60	0.69	0.75	0.78	**0.81**	0.83
2 LTs	0.60	0.74	**0.81**	0.84	0.87	0.88	0.89
3 LTs	0.69	**0.80**	0.85	0.88	0.90	0.91	0.92
4 LTs	0.74	0.84	0.88	0.90	0.92	0.92	0.93
5 LTs	0.77	0.86	0.90	0.91	0.93	0.93	0.94
6 LTs	**0.80**	0.88	0.91	0.92	0.93	0.94	0.94
7 LTs	0.82	0.89	0.92	0.93	0.94	0.94	0.95

Note. Results shown for raters nested within teachers with one rater for each teacher. Bolded results show the threshold of combinations that result in at least learning targets to achieve a generalizability coefficient of at least .80. D-Study = decision study; LT = learning target.

learning targets per teacher, which, in turn, reduces the influence of two of the larger variance components (the learning target and error variance components) on measurement error variance.

Discussion

Following a systematic process for the development, piloting, and use of the LTRS to measure the quality of teacher-developed learning targets for embedded instruction for early learning, we used G-theory to examine sources of variance in LTRS scores and to explore conditions for score dependability in future studies. Results of the G-study showed a large proportion of variance was due to learning targets and teachers, while a small proportion of variance was due to raters, children, and error. Results of the D-study showed dependable LTRS total scores when teachers' learning targets were measured under the data collection and scoring design conditions used in the present study (Snyder et al., 2016). The D-study also showed the dependability of scores will likely not be affected by the number of raters if the procedures for training and monitoring raters used in the present study are employed. The total number of learning targets, however, did have a noticeable impact on the dependability of scores. Taken together, estimates obtained in the present study suggest future studies could be designed with a minimum of one rater (under similar training conditions) and six total learning targets across children in a classroom would meet acceptable levels of score dependability. It is important to note, however, that the low variance due to raters in the present study might be attributed to the intensive training procedures and ongoing inter-rater agreement checks conducted as part of the study. Use of the LTRS under different rater training and agreement check procedures might not generate the low variance due to raters and could affect the dependability of the obtained scores. Although D-studies such as those reported here are useful in forecasting dependability for future studies, the reliability of scores from the LTRS should be examined with each new administration and use of the measure (Snyder, Lawson, Thompson, Stricklin, & Sexton, 1993; Thompson, 2003).

Our aim with estimating variance components and G coefficients for teachers' dimension scores was to understand better how different facets influenced dimension scores and what impact this has on their score dependability. For example, we hypothesized there might be dimensions in which raters accounted for more variance compared with other dimensions; this turned out not to be the case due to the overall small influence of raters on score dependability. There were differences, however, in whether teacher or learning target accounted for more variance across dimensions, which, in turn, affected dependability scores. For example, the level of dependability for dimension scores was variable and suggests there was room for improvement. Results from the present study will be used to guide revisions to the LTRS. In addition, the results provide important information about influences on quality learning targets that have implications for future research, teacher practice, and PD, including preservice training.

For example, we are interested in exploring further the role of the type of learning target in teachers' LTRS scores, given the large proportion of variance that learning targets accounted for across occasions and select dimension scores. We plan to examine learning targets by the content domain targeted and the type of skill specified. Content domain targeted refers to the extent to which the behavior or skill specified in the learning target is focused on different domains of learning (e.g., language, literacy, numeracy, cognitive, motor, self-help). Skill specified refers to four categories described by Snyder et al. (2015) and Wolery and Hemmeter (2011): (a) dispositions (e.g., persistence), (b) chains of behaviors (e.g., washing hands), (c) discrete responses (e.g., naming colors), and (d) response classes (e.g., imitating peers).

It is plausible that the content domain targeted and type of skill indicated in a learning target had more influence on scoring indicators for dimensions such as functionality and generality. This may provide a partial explanation for why learning targets accounted for relatively large percentages of these dimension score variances in the G-study. For example, if a learning target is focused on naming different shapes (preacademic discrete skill) credit would be given for aligning with the general preschool curriculum but credit might not be given for promoting independence or multiple forms of a behavior that serves the same function. Alternatively, if a learning target is focused on using a two to three word phrase to respond to questions (expressive language response class), credit would be given for aligning with the general preschool curriculum and for a skill that promotes independence and allows for multiple response forms.

With respect to the other LTRS dimensions, findings suggest that if a teacher adds a listing of activities or a criterion statement to one type of learning target, he or she can likely add these features to other learning targets, regardless of the content targeted or type of skill. This interpretation helps explain why measurability and instructional context dimensions had large percentages of variance associated with the teacher.

Examining if there are systematic differences in the numbers and types of LTRS indicators marked as present by domain content and type of skill targeted might reveal tensions between specific learning target behaviors and the indicators of quality on the LTRS. Currently, indicators are viewed as applicable across all learning target behaviors with only two not applicable (N/A) conditions specified. Based on the findings from the present study, we intend to review the decision rules for N/A and related indicators to

determine if they need further refinement. Some indicators might be applicable across all learning behaviors regardless of content targeted or skill type (e.g., indicators for behavior statement, age appropriate, and measurable), while indictors associated with functionality, generality, and instructional context might be more or less relevant depending on the content targeted or skill type specified. Additional research and use in practice with the LTRS is needed to analyze these different aspects of the dimensions and scoring rules, including the N/A scoring. For example, in future uses of the LTRS, a rater might first classify the content domain and type of skill and then apply the relevant dimensions and indicators given the learning target categorization.

Examining the dimensions of the LTRS in relation to the facets that influence variance in scores (or teachers' abilities to write quality learning targets) can also provide information to inform PD and teacher training. Data from the present study suggest that variance in teacher scores was greatest for instructional context and measurability such that teachers either include or do not include statements that identify activities for instruction (when to teach) or criteria for evaluating learning (how to evaluate). This suggests some teachers might need further training and support to understand the importance of these practices in supporting learning and their inclusion in a written learning target. In the context of the potential efficacy trial from which LTRS data were used for the present study, this finding might be related to the PD received by participants. Those who participated in the PD received training and follow-up resource support materials (e.g., handouts with definitions and examples, self-assessment checklists) focused on these indicators, while those who received no PD did not.

The potential influences between types of learning targets and LTRS indicators described above suggest teachers need a nuanced understanding of what matters and when for writing learning targets. To achieve this nuanced understanding, varying the type and intensity of professional supports might be needed (Buysse, & Hollingsworth, 2009; Snyder, Hemmeter, & McLaughlin, 2011). For example, early PD supports and learning for teachers in training might be less intensive and focus on a generalized introduction to key indicators of learning target or IEP goal quality (e.g., overview presentations and discussions, handouts with definitions and examples). Over time, PD and learning support might become more complex and embedded in practice with more complex and varied exemplars, more practice, self-assessment checklists with LTRS dimensions included on them, and descriptive performance feedback.

In their study of a training program to improve individualized family service plan (IFSP)/IEP goals for routines-based intervention, Boavida, Aguiar, and McWilliam (2014) worked with 35 early intervention teams for 24 hr across five training sessions using presentations, case exemplars, demonstrations, role-play, and group work followed by 3 months of fieldwork with regular supports and written feedback on the quality of child learning goals through an e-learning platform. These professional supports were effective for improving the quality of written goals. The authors highlighted two implications from their findings, which are also relevant for the present study. First, intensive and varied training is needed to improve practitioners' ability to write quality goals. Second, the ability to write high-quality goals (or learning targets in the case of embedded instruction) is only useful if practitioners are able to use the goals to improve their instruction and outcomes for children.

Individualization of learning targets is a hallmark of the what to teach component of embedded instruction. Data from the G-study showed little variance of children's learning targets when nested within teachers. This finding suggests either teachers specify learning targets for different children that are equally easy or hard to write in relation to LTRS indicators or the nature of teacher-developed learning targets is relatively homogeneous across children. Our review of the learning targets in the present study sample revealed that some teachers used identical or similar skills, conditions, and criteria across the two or three participating children from their classroom. Given the LTRS does not explicitly evaluate individualization, we were not able to determine if similar learning targets were equally appropriate for the different children associated with a teacher. Nonetheless, this finding is consistent with previous research that has identified concerns with individualized programming in inclusive settings (Christle & Yell, 2010; Epsin, Deno, & Albayrak-Kaymak, 1998) and suggests teachers might need more support to individualize priority-learning targets for children. This includes not only the identification of key skills but also skills in relation to the child's phase of learning, conditions of support needed, and appropriate criteria for evaluation. PD focused on improving the quality of instructional learning targets needs to extend beyond what is measured on instruments used in the context of research, such as the LTRS. For example, supplemental materials for the LTRS might include case exemplars and indicators for individualization based on a child's phase of learning and the types of skills or behaviors targeted for embedded instruction.

Considerations for Future Research and Measure Development

Observational or judgment-based rating instruments, such as the LTRS, are being increasingly used in early care and education, including in early childhood special education, to support teachers' implementation of interactional and instructional practices. The use of these instruments in

research contexts requires careful examination of the content of the instrument; the purpose, context, and population for which the measure is intended; recruiting and training of raters; data collection procedures, including sampling of observations or events to be evaluated; scoring; and the interactions among these conditions to ensure dependable scores (Hill, Charalambous, & Kraft, 2012). Moreover, pilot testing and systematically evaluating the psychometric integrity of scores from practice-focused measures is essential (Thompson, 2003).

Results from the present G-study showed the object of measurement (i.e., teachers) did not have the largest observed score variances relative to other facets of measurement. This was also found in a study conducted by Snyder, Hemmeter, Fox, Bishop, and Miller (2013). These authors conducted a G-theory study of the Teaching Pyramid Observation Tool–Pilot Version (T-POT-P; Fox, Hemmeter, & Snyder, 2008). In this study, indicators also accounted for more variance than teachers. Taken together, both studies suggest that variance for some facets of measurement might exceed the variance for the object of measurement. This finding might be expected and acceptable in applied contexts and should not necessarily negate the utility of scores obtained. As measures of effective teacher interactions or instruction using observations or ratings of practice implementation increase in use, careful consideration of the influences on teachers' scores and the implications for score interpretation in the context of applied settings is warranted.

Considerations for Embedded Instruction Practice

Given an increased focus on intentional teaching and use of systematic interventions like embedded instruction to promote child learning and engagement in inclusive early learning settings (DEC, 2014), a need exists to support preschool teachers to use embedded instruction practices with fidelity. The embedded instruction practice of interest in this study was teachers' ability to write quality learning targets to inform the what to teach component of embedded instruction practices. The use of the LTRS under the conditions of administration in the present study might not be feasible or desirable in practice. Nonetheless, indicators and dimensions on the LTRS might offer preschool teachers and those who support PD a useful tool for developing quality learning targets for embedded instruction. As described previously, additional indicators or adaptation to existing indicators might be needed when the LTRS is used in practice contexts. Nevertheless, we assert use of measures such as the LTRS are important for at least two reasons. First, the quality of written priority-learning statements varies across states, districts, and teachers (Christle & Yell, 2010; Huefner, 2000). Resources and PD materials might help teachers and other related services personnel ensure learning targets are of high quality and thus more likely to support effective instruction. Second, embedded instruction involves the specification of what to teach, when to teach, how to teach, and how to evaluate teaching (Snyder, Hemmeter, Sandall, McLean, & McLaughlin, 2013). These components are interrelated and influence the overall quality of the embedded learning opportunities for each child. A clearly defined learning target is viewed as a necessary component for ensuring intentional teaching episodes occur in the context of everyday activities. If teachers and other related services personnel want to examine, observe, quantify, or provide feedback on the dynamic aspects of instruction, such as teacher implementation of embedded instruction or measuring child progress, a clearly defined learning target is necessary (Boavida et al., 2014; VanDerHeyden, Snyder, Smith, Sevin, & Longwell, 2005). To illustrate this point, we provide learning targets that have been evaluated as low quality and high quality using the LTRS. As shown in the appendix, information included in a high-quality target (i.e., a target that addresses the key dimensions and indicators) informs what, when, and how embedded instruction should be implemented as well as how instruction should be evaluated.

Measures such as the LTRS, which are designed to assess teachers' implementation of teaching practices in the context of research projects, might also contribute to supporting practitioners' understanding of key practices associated with multi-component interventions such as embedded instruction. Future research might explore how understanding and application of the key dimensions of the LTRS are interpreted as it is adapted for use from a research to a practice context.

Conclusion

The present study provides information about the development and design of the LTRS and the conditions most likely to result in dependable scores for future use in research and practice. Results illustrate the utility of generalizability theory to design studies with adequate score dependability for measures of teachers' implementation of practices, particularly those used to examine intervention effects in experimental studies or assess teacher performance (Hill, Charalambous, Blazer, et al., 2012; Hill, Charalambous, & Kraft, 2012).

A G-theory study provides a behind-the-scenes look at what is influencing score dependability. These types of studies typically are used to evaluate and refine a measurement instrument. In the present study, we have indicated how data might also be used to understand key aspects of the constructs and associated indicators being measured and implications of the findings for practice, including for professional learning and development.

Appendix

Low-Quality Versus High-Quality Learning Targets as Measured by Number of LTRS Indicators Present.

Learning target	Indicators present[a]	LTRS score	Implications for embedded instruction
Sean will recognize numerals 1 to 10.	Behavior specified Written clearly Age appropriate skill Preschool curriculum Independence Embeddable Natural environment	50%	This learning target begins to identify **what to teach**, but use of the term "recognize" does not specify how Sean will demonstrate the target behavior. Related to what to teach, the skill selected aligns with the preschool curriculum, is age appropriate, and numeracy skills can support independence. Related to **when to teach**, it is the type of skill that can be embedded in a range of different activities throughout the day and the learning target does not indicate a restrictive setting for instruction, yet the information about when to teach is not sufficient. In addition, no information about **how to teach** or **how to evaluate** is included.
Sean will point to or give the numerals 1 to 10 (e.g., show me the number 9) with verbal cues.	Above plus + Demonstration specified Exemplars provided Multiple forms Observable Conditions listed	86%	In this learning target, the ways in which Sean might "recognize" a numeral such as pointing to or giving numerals has been included, making the **what to teach** information clearer for the teaching team. The learning target also includes information about **how to teach** by providing examples of ways to create teaching opportunities and the conditions of help to support Sean (i.e., verbal cues). Additional information to clarify **when to teach** or **how to evaluate** has not been added.
Sean will point to or give the numerals 1 to 10 (e.g., show me the number 9) with verbal cues across different activities including small group, centers, and outdoor play. I will know he can do this when he recognizes each numeral without correction on two different occasions.	Above plus + Across activities Criterion stated	100%	In this learning target, previous information was sufficient for supporting the **what to teach** and **how to teach** components of embedded instruction. With the inclusion of activities for embedding instruction on the target included here, there is sufficient information about **when to teach**. The addition of a criterion statement has also provided important information for **how to evaluate** Sean's learning.

Note. The example learning targets are provided to illustrate the types of learning targets preschool teachers might write, how they vary in quality as measured by the LTRS, and the implications for how different learning targets can inform embedded instruction. LTRS = Learning Target Rating Scale.
[a]For this type of learning target, both materials and interactions with peers would be scored not applicable (n/a).

Authors' Note

The opinions expressed are those of the authors, not the funding agency. Tara McLaughlin was previously at the University of Florida and is still associated with the Anita Zucker Center for Excellence in Early Childhood Studies at the University of Florida.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Work completed in this article was supported, in part, by a grant from the National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education to the University of Florida (R324A070008).

References

Boavida, T., Aguiar, C., & McWilliam, R. A. (2014). A training program to improve IFSP/IEP goals and objectives through the routines-based interview. *Topics in Early Childhood Special Education*, *33*, 200–211.

Buysse, V., & Hollingsworth, H. L. (2009). Program quality and early childhood inclusion: Recommendations for professional development. *Topics in Early Childhood Special Education*, *29*, 119–128.

Christle, C., & Yell, M. (2010). Individualized education programs: Legal requirements and research findings. *Exceptionality: A Special Education Journal*, *18*, 109–123.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Mason, OH: Thomson Wadsworth.

Division for Early Childhood. (2014). *DEC recommended practices in early intervention/early childhood special education*. Retrieved from http://www.dec-sped.org/dec-recommended-practices

Drasgow, E., Yell, M. L., & Robinson, R. (2001). Developing legally correct and educationally appropriate IEPs. *Remedial and Special Education*, *22*, 359–373.

Epsin, C. A., Deno, S. L., & Albayrak-Kaymak, D. (1998). Individualized education programs in resource and inclusive settings: How "individualized" are they? *The Journal of Special Education*, *32*, 164–174.

Fox, L., Hemmeter, M. L., & Snyder, P. (2008). *The Teaching Pyramid Observation Tool Research Edition* (Unpublished assessment). Tampa: University of South Florida.

Giangreco, M. F., Dennis, R. E., Edelman, S. W., & Cloninger, C. J. (1994). Dressing your IEPs for the general education climate: Analysis of IEP goals and objectives for students with multiple disabilities. *Remedial and Special Education*, *15*, 288–296.

Goodman, J. F., & Bond, L. (1993). The individualized education program: A retrospective critique. *The Journal of Special Education*, *26*, 408–422.

Grisham-Brown, J., & Hemmeter, M. L. (1998). Writing IEP goals and objectives: Reflecting an activity-based approach to instruction for young children with disabilities. *Young Exceptional Children*, *1*(3), 2–10.

Hill, H. C., Charalambous, C. Y., Blazer, D., McGinn, D., Kraft, M. A., Beisiegel, M., . . . Lynch, K. (2012). Validating arguments for observational instruments: Attending to multiple sources of variation. *Educational Assessment*, *17*, 88–106. doi:10.1080/10627197.2012.715019

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*, 56–64. doi:10.3102/0013189X12437203

Horn, E., Lieber, J., Li, S., Sandall, S., & Schwartz, I. (2000). Supporting young children's IEP goals in inclusive settings through embedded learning opportunities. *Topics in Early Childhood and Special Education*, *20*, 208–223.

Huefner, D. S. (2000). The risks and opportunities of the IEP requirements under IDEA '97. *The Journal of Special Education*, *33*, 195–204.

Hunt, P., Goetz, L., & Anderson, J. (1986). The quality of IEP objectives associated with placement on integrated versus segregated school sites. *The Journal of the Association for Persons with Severe Handicaps*, *11*, 125–130.

Lynch, E. C., & Beare, P. L. (1990). The quality of IEP objectives and their relevance to instruction for students with mental retardation and behavioral disorders. *Remedial and Special Education*, *11*, 48–55.

McLaughlin, T., & Snyder, P. (2014). Embedded instruction to enhance social-emotional skills. In J. Hart & K. Whalon (Eds.), *Friendship 101: Helping students build social competence* (pp. 63–78). Arlington, VA: Council for Exceptional Children.

Michnowicz, L., McConnell, S., Peterson, C., & Odom, S. (1995). Social goals and objectives of preschool IEPs: A content analysis. *Journal of Early Intervention*, *19*, 273–282.

Notari, A. R., & Bricker, D. D. (1990). The utility of a curriculum-based assessment instrument in the development of individualized education plans for infants and young children. *Journal of Early Intervention*, *14*, 117–132.

Pretti-Frontczak, K., & Bricker, D. (2000). Enhancing the quality of individualized education plan (IEP) goals and objectives. *Journal of Early Intervention*, *23*, 92–105.

Semmelroth, C. L., & Johnson, E. (2014). Measuring rater reliability on a special education observation tool. *Assessment for Effective Intervention*, *39*, 131–145.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: SAGE.

Snyder, P., Hemmeter, M. L., Fox, L., Bishop, C., & Miller, D. (2013). Developing and gathering psychometric evidence for a fidelity instrument: The Teaching Pyramid Observation Tool–Pilot version. *Journal of Early Intervention*, *35*, 150–172.

Snyder, P., Hemmeter, M. L., & McLaughlin, T. (2011). Professional development in early childhood intervention: Where we stand on the silver anniversary of P.L. 99-457. *Journal of Early Intervention*, *33*, 357–370. doi:10.1177/1053815111428336

Snyder, P., Hemmeter, M. L., McLaughlin, T., Algina, J., Sandall, S., & McLean, M. (2011, April). *Impact of professional development on preschool teachers' use of embedded-instruction practices*. Paper presented for the American Educational Research Association Annual Conference, New Orleans, LA.

Snyder, P., Hemmeter, M. L., McLean, M., Sandall, S., McLaughlin, T., & Algina, J. (2016). *Impact of professional development on preschool teachers' use of embedded-instruction practices*. Manuscript in preparation.

Snyder, P., Hemmeter, M. L., Sandall, S., McLean, M., & McLaughlin, T. (2013). Embedded instruction practices in the context of response to intervention. In V. Buysse & E. Peisner-Feinberg (Eds.), *Handbook of response-to-intervention in early childhood* (pp. 283–300). Baltimore, MD: Brookes.

Snyder, P., Lawson, S., Thompson, B., Stricklin, S., & Sexton, D. (1993). Evaluating the psychometric integrity of instruments used in early intervention research: The Battelle Developmental Inventory. *Topics in Early Childhood Special Education*, *13*, 216–232.

Snyder, P., McLaughlin, T., Sandall, S., McLean, M., Hemmeter, M. L., Crow, R., . . . Embedded Instruction for Early Learning Project. (2008). *LTRS: Learning target rating scale: Research version 1 [Manual]* (Unpublished instrument, IES R324A070008).

Snyder, P., McLaughlin, T., Sandall, S., McLean, M., Hemmeter, M. L., Crow, R., . . . Embedded Instruction for Early Learning Project. (2009). *LTRS: Learning target rating scale: Research version 2 [Manual]* (Unpublished instrument, IES R324A070008).

Snyder, P., Rakap, S., Hemmeter, M. L., McLaughlin, T., Sandall, S., & McLean, M. (2015). Naturalistic instructional approaches in early learning: A systematic review. *Journal of Early Intervention*, *37*, 69–97.

Thompson, B. (2003). *Score reliability: Contemporary thinking on reliability issues*. Thousand Oaks, CA: SAGE.

VanDerHeyden, A. M., Snyder, P., Smith, A., Sevin, B., & Longwell, J. (2005). Effects of complete learning trials on child engagement. *Topics in Early Childhood Special Education*, *25*, 81–94.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2007). Reliability coefficients and generalizability theory. In C. R. Rao (Ed.), *Handbook of Statistics: Vol. 26. Volume on psychometrics* (pp. 81–124). London, England: Elsevier.

Wolery, M., & Hemmeter, M. L. (2011). Classroom instruction: Background, assumptions, and challenges. *Journal of Early Intervention*, *33*, 371–380.

Zaslow, M., Martinez-Beck, I., Tout, K., Halle, T., Ginsberg, H., & Hyson, M. L. (Eds.). (2011). *Quality measures in early childhood settings*. Baltimore, MD: Brookes.